PCT/GB 2004 / 0 0 1 3 9

PATENTS • DESIGNS • The Patent Office COPYRIGHT • TRADE MARKS

INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed

Dated 2 August 2004

BEST AVAILABLE COPY

For Official use only

Your reference Prob Matching (UK)

0308413.4 11 APR 2003

The Patent Office

Request for grant of a Patent

Form 1/77 Patents Act 1977

1 Title of invention

Verification of authenticity of check data

2. Applicant's details

☒ First or only applicant

2a If applying as a corporate body: Corporate Name

EnSeal Systems Limited

Country

GB

2b If applying as an individual or partnership

Surname

Forenames

2c Address

6 Thorney Leys Business Park
Witney
Oxford

UK Postcode OX8 7GE

Country GB

ADP Number 8321291001

□ Second applicant (if any)

2d Corporate Name

Country

2e Surname

Forenames

2f Address

UK Postcode

Country

ADP Number

3 Address for service

Agent's Name	Origin Limited
Agent's Address	52 Muswell Hill Road London
Agent's postcode	N10 3JR
Agent's ADP Number	C03274

7270457002

4 Reference Number

Prob Matching (UK)

5 Claiming an earlier application date

An earlier filing date is claimed:

Yes ☐ No ☒

Number of earlier application or patent number

Filing date

15 (4) (Divisional) ☐ 8(3) ☐ 12(6) ☐ 37(4) ☐

6 Declaration of priority

Country of filing	Priority Application Number	Filing Date

7 Inventorship

The applicant(s) are the sole inventors/joint inventors

Yes □ No ☒

8 Checklist

		Continuation sheets	0
Claims	2	Description	13
Abstract	0	Drawings	1 only

CF

Priority Documents ~~Yes~~/[✓ No]

Translations of Priority Documents ~~Yes~~/[✓ No]

Patents Form 7/77 ~~Yes~~/[✓ No]

Patents Form 9/77 [✓ Yes]/~~No~~

Patents Form 10/77 ~~Yes~~/[✓ No]

9 Request

We request the grant of a patent on the basis of this application

Signed: Origin Limited Date: 11 April 2003

(Origin Limited)

VERIFICATION OF AUTHENTICITY OF CHECK DATA

Overview

This invention concerns the automatic detection of data falsification on printed checks using high speed, low resolution scanners as the means of image acquisition prior to analysis. As with the prior art, the basis of the method is the comparison of human readable information with machine readable information added to the check at the time of printing.

The improvement which constitutes the major part of the invention arises from the recognition that both human readable data and the machine readable data are subject to error during the printing and scanning processes and a mismatch in the two forms of data may not necessarily imply fraudulent alteration. In practical terms the form of coding for the machine readable data is made to depend upon the characteristics of the human readable text and its retrievability and consists of independent segments that allow for partial recovery despite localised degradation The analyses of the text and machine readable code are mutually dependent and together with external data provide a probability model for the detection of possible fraudulent checks.

Background

There is a need to provide a cheap and rapid means of corroborating the authenticity of the critical, human readable data on checks in order to identify fraudulent falsification. Checks are the subject of high speed printing and scanning operations: operational constraints generally require anti fraud techniques to integrate with existing schemes. Thus a number of methods have been proposed in which additional symbols or data are printed onto checks, the same data being subsequently scanned and analysed by image sorters.

The problem with methods that have so far been proposed is that no proper account is taken of the degradation that may well occur to the added symbols as well as the inevitable misreading that is inherent in OCR. Simply rejecting checks where the OCR does not match the retrieved machine readable data would result in large numbers of satisfactory checks being sent for inspection. In most methods the machine encoded data is some representation of the totality of all of the data so that damage to a part of the representation removes the possibility of any meaningful data retrieval.

Verification of checks by adding machine readable symbols has a long history. A method of authenticating check data was provided by Szepenski (German Patent 29 43 436 A1) in 1979, although his description was not particularly concerned with the workflow issues associated with image sorters and the like. Text on documents in his method was to be authenticated by means of a machine readable pattern which contained the same information as the human readable text and extended over the whole document. The pattern was to contain all of the textual information and in a paper published more or less concurrently Szepenski suggested the use of standard error correction techniques to overcome the inevitable problems of accurate machine reading.

In 1994 EP 0 699 327 B1 Abathorn issued a patent which described a modified version of Szepenski's method in which the machine readable data is in the form of a bar code (or other symbology which is not specified) to be added to the check. This patent goes further by describing how the data might be added " in a single pass through the printing system enabling high speed automated mass production of bearer documents." There is little description of the coding method but the fact that "if a user's name has been obscured, the name can be recovered if the name was selected as a value critical data item" suggests that the data is not hashed or encrypted.

Ramzy, USPTO 6,073,121 also describes a method of protecting a check by adding machine readable data, in this case the data comprising " all the check data" in the form of a bar code or other symbol. Ramzy differs from Abathorn in that the added data is encrypted. The implication is that the data retrieved from the bar code must be retrieved in its entirety or else it is not decipherable, and this implies that the bar code or other symbol must be robust against poor quality imaging.

In USPTO 6,243,480 Zhao et al authenticate check data by adding "authentication information" in machine readable form, this form either being a watermark or a symbol which could be a bar code. The authentication information described in the patent is some form of digest of semantic information. The digest formed from the OCR allows a certain amount of latitude in that commonly confused characters such as "c" and "e" are allowed to be in error without destroying the correspondence between the two versions of the authentication information. However, the machine readable code is such that corruption of it is not reversible and there is no possibility of relaxing the equality condition if, for instance, a bar code is damaged by a scanning problem. As is stated in the claims " an authentication information reader reads the first authentication information ....." and compares with data from "an authenticator that computes the second authentication information. " "Reading" the information as opposed to computing information in general allows no scope for adjusting to poor quality images.

Similarly in USPTO 6,170,744 Payformance tackles the problem of self authentication by including authenticating data in machine readable form. The authenticating data as above includes some form of digest in the form of a hash, signature or encryption and in each case the data is not reversible or would not be reversible if some uncorrectable reading error occurred. The verification is by equality of two values and has no provision for close misses or data adjustment.

In USPTO 6,233,340 Sandru describe yet another method of adding authenticating data and here again the data is concatenated in some way which prevents its being deciphered when damaged by the imaging process.

Summary

As described above, methods hitherto advocated for authentication of documents include the comparison of data printed in at least two different forms on the document. The documents are scanned at the time of authentication and the images are analysed to allow the comparison to be made. Each form of data appearing on the document will require its own algorithm to retrieve the encoded data. This algorithm may be a form of OCR, or a bar code interpreter, or a customised interpreter for special forms of encoding such as the Seal encoding which is part of one implementation of the present invention.

The data that is added usually originates in the form of a string of alphanumeric characters that may be part or whole of the data on the document. In the case of checks the data that is embedded could be any selection of the variable data as opposed to the check stock data. This data includes payee, amount, account number, date, bank routing number and data unique to a particular bank.

In the traditional printing of checks the added data simply appears in text form and this will also be the case in the implementation of this invention. Thus the data consists of a set of distinguishable characters. This data or a subset or digest of this data is added in a form that is machine readable and generally not human readable. The important distinction in methods described in the prior art is between those that aggregate the characters in some manner and calculate a representative value and those that encode the characters individually. The implication is that where data has been aggregated any failure in the retrieval process may render the whole of the data invalid, whereas if the data is segregated damage to parts of the data may leave the remaining data decipherable.

The two commonly used forms of data aggregation are encryption and hashing. In all standard encryption algorithms, e.g DES, RSA, Blowfish, it is regarded as important that each bit of the plaintext affects every other bit to produce the ciphertext, this requirement rendering the breaking of the code much more difficult. A consequence of this is that alteration of any portion of the cipher text has a potential effect on every bit of the plaintext.

Thus if ciphertext is embedded in the machine readable code and any part of that code cannot be correctly retrieved the whole of the plaintext is invalid.

In the case of hashing, a similar situation holds. Hashing algorithms e.g MD5, SHA1 etc are designed so that hashed values which differ slightly correspond to originals that differ considerably. Again, if a hash value is embedded in machine readable form and that hash value is corrupted, however slightly, by the retrieval process, then the original data cannot be reconstructed.

Thus in those versions of the prior art which use encryption or hashing, any misread in either OCR or the machine readable data will result in mismatch of the values that are required for authentication. The only outcome of such a comparison is agreement or non agreement and the level of disagreement is identical whether one or all of the original data characters is misread by the OCR, or whether one or all of the bits of the version of the hash value after error correction is altered. Given that for the data on checks the probability of correct identification of all of the human readable characters is at best 98 to 99% then a huge number of checks will be incorrectly identified as fraudulent.

It is one of the objects of this invention to remedy this situation by embedding data in machine readable form in discrete segments so that if one segment is damaged the remainder may still be valid and able to give information about likelihood of deliberate falsification.

It is also well known that certain characters are easily confused by the OCR process. characters such as O and 0, C and O, F and E etc. Now in USPTO 6,243,480(Mediasec) these sorts of characters are allowed to be considered interchangeable to reduce the OCR reading errors, but no information about how much confusion has occurred will be available.

In the present invention actual individual coding of these characters is such that their chance of confusion in the machine readable code is inversely proportional to their chance of confusion by the OCR methods. That is to say one form of data embedding is a function of the retrieval probabilities of another form of data embedding.

It is a further object of this invention to develop a function which predicts the probability of deliberate falsification, as opposed to misreading, by constructing the data retrieval process to return information about the nature of any errors. Thus the probability function will be a function of the measured quality of the image machine readable code and the human readable data, measured by the fact that these entities give clear, unambiguous symbols or are difficult to resolve. The probability will also be a function of such parameters as the relative position of erroneously detected characters, having regard to the fact that falsification usually involves a coherent set of contiguous characters rather than randomly separated characters.

Detail of Method

Workflow

Many large corporations print their own checks in a bulk processing environment using high speed printers, usually laser printers. The usual method is to have check stock preprinted with general information about the Bank, its routing number and similar data which is common to thousands of checks. The individualised data required before issuance of the check includes payee name, account number, date, amount of transaction etc. and this is usually added by a laser printer.

In the present invention a seal or other machine readable code is printed at the same time as the variable data is added. This is generally achieved by adding an image of the seal to the printing file before it is despatched to the printer, but it can be equally well achieved by modifying the PCL commands or the use of soft fonts if they are the means that will best accord with the running of the system.

In the normal cycle of the check the payee pays in the check to the "Bank of First Deposit" or to a check cashing outlet. At this point the human readable data is read and possibly a cash payment takes place. In one implementation of this invention the Seal containing authenticating data will also be read using a simple desk top scanner or equivalent check reader.

The check is then forwarded to the issuing Bank or financial services company acting for the Bank. High speed scanners are used to capture images of both the front and back of the check in a bulk processing mode with minimal human intervention. The analysis of data and reconciliation of the checks then takes place using the images. In accordance with this invention the data on any Seals will be read and analysed either this point or as part of an offline process. At this point also any checks which do not meet acceptance criteria, perhaps on account of being damaged or unreadable or because two forms of data do not match, will be identified as exceptions and subjected to further examination.

Data for Machine Readable Code

The variable data that is printed on the checks just prior to issuance includes the payee name, the amount, the account number, the date. The amount and account number are also included in the MICR line printed at the bottom of the check and which provides another machine readable source of data.

In the proposed implementations a seal containing at least two of these entities in machine readable form will be printed onto the check.

The data is in the form of alphanumeric characters which are converted to binary strings before being represented in graphical form. An important feature of the conversion to binary format is the fact that the string consists of independent but interleaved segments, each segment representing a character or small group of characters. Thus if 10 letters were to be converted to a binary string, each letter might be represented by a16 bits and these bits might be interspersed in a string of 160 bits according to some rule. If one of the letters were to be changed only the 16 corresponding bits would be changed and there would be no knock on effect on the rest. Similarly, if 1 bit were to be changed only one of the letters would be affected.

The manner of representation of characters in binary form is a key part of this invention. In many applications the codes representing characters are generated using an established error coding technique. Often used are cyclic codes on account of their structure which lends itself to easy decoding. In the case of this invention there is no need for highly structured codes because the chunks of data to be decoded are small enough to be handled by cruder methods. The main requirement is that the "Hamming Distances " (HD) between codes should be chosen so as to best reflect the quality of information derivable from the scanned images.

The HD between two codes of equal bit length is simply the number of bit positions in which the codes differ. Thus if 2 codes have a large HD they are unlikely to be confused unless there is a large number of bit errors. The penalty for making HD's too large is that the codes become too long and occupy too much of the available payload.

The factors affecting the HD's are, according to this invention:

(a) The quality of the images of the seals.

In most implementations there will be many checks available to standardise data and find expected values for any quality measurements.

The quality of the images is a function of the resolution of the scanners, their quality in terms of tendency to merge distinct features or produce artefacts and any issues arising from the rapid transport of checks through the processing system.

The quality is also a function of the consistency of the printing method and such matters as level of toner within a printer.

This quality has to determine the overall distribution of HD's of any set of codes, ensuring that the likelihood of a mistead is at a satisfactory level. Thus if quality is very poor the number of bits in the codes will be increased to allow a greater error margin.

(b) The accuracy of any OCR reader

The number of errors produced by the OCR reader should give an additional guide to the accuracy required of a Seal and hence the overall distribution of HD's.

In addition HD's should be adjusted to take into account the fact that some characters are far more likely to be confused by OCR than others. "O" for instance is frequently mistaken for "C" but "Z" is rarely mistaken for "I."
To cope with this property of OCR the HD's between the code for "O" and code for "C" will tend to be larger than between those for "Z" and "I." Thus although the OCR may tend to confuse "O" and "C" the Seal reading would be highly unlikely to do so.

With these considerations in mind a set of codes can be generated to represent the characters and hence convert the human readable text into a binary string.

Representation of Data in Machine Readable Form

In a preferred embodiment the form in which the data is added is that described in detail in the Bitmorph patent PCT/GB02/00539.

In an alternative embodiment the data is added in the form of a two dimensional bar code.

Analysis of Seal

The scanners provide images of checks, generally in black white, for the purposes of analysis. A further source of data may be from the reading of the MICR line by a device which reads magnetic ink.

Where there is machine readable code such as that produced by bar codes, glyphs or Seals there are many well described techniques to orientate and scale images prior to analysis of individual code bearing symbols. For the purposes of this description it will be assumed that the analysis can be taken to the level where the information is contained in a set of graphics, each graphic being a cell containing a configuration of black and white pixels which is to be interpreted.

Thus where glyphs are used the cells will typically be squares containing black pixels which in the original image formed a diagonal stripe, the orientation of the stripe indicating whether the symbol is to be counted as a "1" or a "0." This configuration will be modified by the printing and scanning processes so that what was originally a sharp clear line will become a more irregular feature. The task of the decoder is to interpret whether such a feature was meant to be a forward or backward sloping diagonal.

Similarly if a Seal is used the cells will be of a variety of shapes and will contain configurations that may originally vertical or horizontal lines but in the scanned images will appear as more diffuse shapes.

In two dimensional bar codes the cells will typically be rectangles each containing 4 black rectangular segments and 4 white spaces in the original form, but after scanning containing irregularities.

It is one of the purposes of this invention to assess any of these forms of machine readable code for the level of degradation and provide a representative quality statistic. By analysing the distribution of this statistic for a large number of checks and associating the quality with the number of errors that is produced in the corresponding decoding process a prediction of likely errors for a given image may be produced.

For glyphs and Seals a set of graphics will correspond to a binary string representing a single character. For instance, each of 40 characters may be represented by 16 bits with HD's chosen appropriately, in other words 16 graphics go to make up a single character. The analysis will allocate to each graphic a "1" or "0" to correspond to the binary string. In many cases there will be several of the bits interpreted wrongly. If the number of errors is within the bounds that the error correction can rectify, the character that is allocated will be that whose binary string has the smallest HD from the interpreted graphics.

In some implementations instead of allocating one of two possible values to a graphic, range of values will be allocated .A number 100 might indicate , for example, a perfect vertical stripe, whilst –100 might indicate a perfect horizontal stripe. A value of +50 would correspond to a vertical stripe with some extra artefacts. Fig.1 shows a set of graphics before and after scanning with a set of values allocated according to the closeness to a vertical or horizontal stripe.

Calculation of HD is modified thus. A binary code such as 1110 0011 1100 1110 is allocated 16 values by replacing each "1" by "100" and replacing each "0" by –100.

Thus the code becomes

{100,100,100,-100, -100,-100,100,100, 100,100,-100,-100, 100,100,100,-100}

The set of scanned graphics corresponding to a character might become, for example,

{ 80, 70, 70,- 20 , -30, -50, -10, -20, 70, 90,,-90,-50 50, 60, 50,. 0}

The HD of this set of 16 graphics from the code would be the sum of the differences for each of the 16 components. That is,

HD between the scanned code and the tested character

= 20 + 30 + 30 +80 +70 + 50 + 110 +120 + 30 + 10 + 10 + 50 + 50 +40 +50 +100

= 850.

The same calculation would be carried out for each of the codes and the code with the smallest HD would be code presumed to correspond to the original data.

Each set of graphics will be tested against the chosen vocabulary of characters. In each case there will be an adjustment (corresponding to the value 850 above) needed to match a given scanned code to one of the vocabulary codes. The sum of the adjustments gives another metric for comparing the quality of the scanned image.

The calculation just described is a non-limiting example of a further aspect of the invention. The decoding of the Seal gives a most probable set of values for the characters. In addition the decoding of the seal allows the allocation of probabilities to one character rather than another. Thus if for a set of 16 graphics the HD from the letter "A" were to be 800 and the HD from the letter "B" were to be 850 there would be quite a high probability that if an "A" appeared where a "B" was expected then this was due to reading error rather than deliberate falsification.

Optical Character Recognition (OCR)

The variable data, in particular the payee name and the amount are read automatically from the scanned images by one of the many available OCR software applications.

In a preferred implementation of this invention the OCR application reads the human readable characters on the check and attributes a probability to some or all of the characters in the selected vocabulary. In general the probabilities are only relevant for two or three characters whose shape most nearly approximates the scanned in figure.

In another implementation the characters that are read from the Seal are passed to the OCR application. The application then considers each supposed character and attributes a probability to the hypothesis that the character read by the OCR is indeed the one proposed by the Seal.

Combining OCR Data and Seal Data

From the foregoing it can be seen that after the Seal reading and the OCR there will be two sets of data which must be compared to authenticate the check in question.

If the OCR data is identical to the Seal data then the check is accepted as authentic.If one or more characters differ then an assessment has to be made as to the cause and the recommended action.

In one implementation the assessment might be as follows.

Firstly a measure is taken of the degree of difference between the OCR data and the Seal data. This might be measured by a metric such as the Levenstein distance which takes into account characters that are substituted, omitted or inserted, or, more appropriately by a metric that is specially tailored to match the known attributes of the system. The metric will include recognition of the close similarity between certain pairs of characters. Thus if a Z appeared where an I were expected a distance of 1.0 might be ascribed, but if a O appeared in place of an 0 a distance of 0.2 might be ascribed.

This metric also takes into account the possible misreads in the Seal where probabilities can be attached through knowledge of the HD's between characters.

Modification of the measured distance can result from assessment of the signficance of the positions in the text in which differences occur. If , for instance, three unmatched characters were randomly distributed through the payee text then it is less likely to be the result of deliberate falsification.

Analysis of the image can be carried out to identify artefacts that have been produced by the scanning process. Such artefacts are often easily recognised as arising from the movement of the check . A further quality factor is the darkness of the image which depends both on the amount of toner added at the time of printing and the threshold value of the scanner.

The extent to which the quality factors affect the Seal and OCR is assessed empirically by sampling large numbers of checks. This sampling will provide an ongoing standardisation.

The overall result is a metric for the difference between Seal and OCR data that is dependent on environmental factors , methods selected for coding and means of interpreting code in graphic form.

In one implementation the MICR information on the check is read and compared with the supposedly identical information in the Seal. The accuracy or otherwise of this comparison is an indicator of the quality of the Seal data, particularly because the MICR information is read to a high degree of accuracy.

Once the difference between Seal data and OCR data has been calculated a threshold has to be decided upon so that checks on one side of the threshold are further examined to see if they might be counterfeit. The level of the threshold depends upon the penalties for false positives and the known likelihood of counterfeits..

Claims

1. A method for assisting the identification of fraudulent checks, the method comprising:-

(a) Encoding algebraically the critical human readable data from the check, where the data is in the form of characters from a known alphabet, converting the algebraic information into a graphical form, printing the graphical form onto the check at the same time as the human readable data is printed;

(b) Scanning the said check;

(c) Reading the human readable information using an OCR scheme which allocates probabilities of each member of the alphabet corresponding to any feature identified as a character;

(d) Reading the data from the encoded graphical form and allocating probabilities of each member of the alphabet corresponding to any feature identified as a character;

(e) Comparing the resulting sets of probabilities and establishing an overall probability that any mismatch is due to reading error rather than deliberate falsification.

2. The method of Claim 1 where the form of coding of the machine readable graphic is dependent upon the characteristics and retrievability of the human readable form.

3. The method of Claim 2 where the Hamming distance between binary representations of a pair of characters will be greatest for those pairs which are least likely to be easily differentiated by an OCR method.

4. The method of Claim 1 where the machine readable code consists of independent segments that enable recovery of partial information when there is localised degradation.

5. The method of claim 1 where the analysis of the graphical form provides a measure of the degradation of the image of the check and this measure in turn and assists in the attribution of probabilities to the likelihood of misreading data.

6 The method of claim 1 where the degradation of the human readable text is assessed by image processing methods and assists in the attribution of probabilities to the likelihood of misreading data.

7 The method of claim 1 where the probability of occurrence of fraud is a known distribution and an algorithm exists which combined with the above probabilities provides a rule for selecting likely exceptions.

8 The method of claim 7 where the probability is assessed with reference to the distribution of errors within the text.

9. The method of claim 1 where the set of elements that make up a character in the representation in graphical form are distributed throughout that form so as to survive moderate localised degradation.

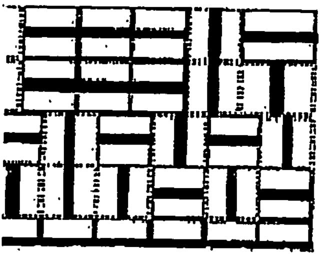

fig. 1(a) Part of Seal 5 x 4 cells
(outlines of cells do not appear in actual Seal)

100 : 100 : 100 :-100 : 100

100 :100 :100 :-100 : 100

-100 : 100 : -100: 100 : -100

-100 : -100: 100 :-100 : -100

fig.1(b) Values Attached to Each Graphic

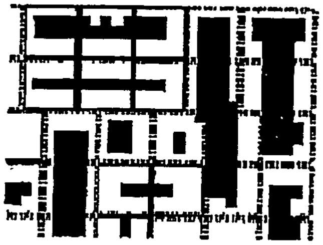

fig. 1(c) Scan of above part Seal
(outlines of cells shown only to indicate method)

90 : 75 : 80 : -80 : 80

90 :100: 85: -60 :- 70

-60: -60: -55: -50; -20

-70: 80; 80; -20; -20

fig.1(d) Values Attached to Each Scanned Graphic